华中科技大学

计算机视觉课程结课报告

文献综述性报告：SAM 图像分割基础模型与相关衍生技术综述

姓　　名：　　李嘉鹏

学　　院：　　计算机科学与技术学院

专　　业：　　数据科学与大数据技术

班　　级：　　大数据 2101 班

学　　号：　　U202115652

指导教师：　　刘康

2023 年 1 月 15 日

目　录

结课报告 SAM 图像分割基础模型与相关衍生技术综述

1 摘要

近年来，人工智能正在向通用人工智能（Artificial General Intelligence，AGI）方向发展。AGI 能够执行广泛的任务，这与狭义上的人工智能模型形成了巨大的对比，因为它们只能完成特定任务，而对其它的下游任务效果不佳。

“基础模型[2]（Foundational Models）”是指在广泛的数据上训练的预训练模型，能够适应各种下游任务。为了提高模型的泛化性能，设计并训练通用的基础模型迫在眉睫。2023 年 4 月 6 日，MetaAI 公布的 SAM 基础模型（Segment Anything Model[1]）突破了传统图像分割的界限，极大地促进了计算机视觉领域基础模型的发展，被视为计算机视觉研究的一个里程碑。

在这一报告中，我将会首先介绍 SAM 的基础架构，然后分析列举基于 SAM 基础模型的相关衍生技术，并讨论其各自的应用领域与特点，最后提出了一些关于 SAM 架构的改进思路。

2 背景介绍

2.1 基础模型概念介绍

基础模型[2]（如 CLIP、SAM、BERT、GPT4 等）在过去几年中已经深刻改变了 AI 领域的发展，这主要是由于它们在大规模数据集上进行了全面的预训练，在广泛的下游任务中具有强大的零样本（zero-shot）泛化能力。目前常见的基础模型如图 1 所示。

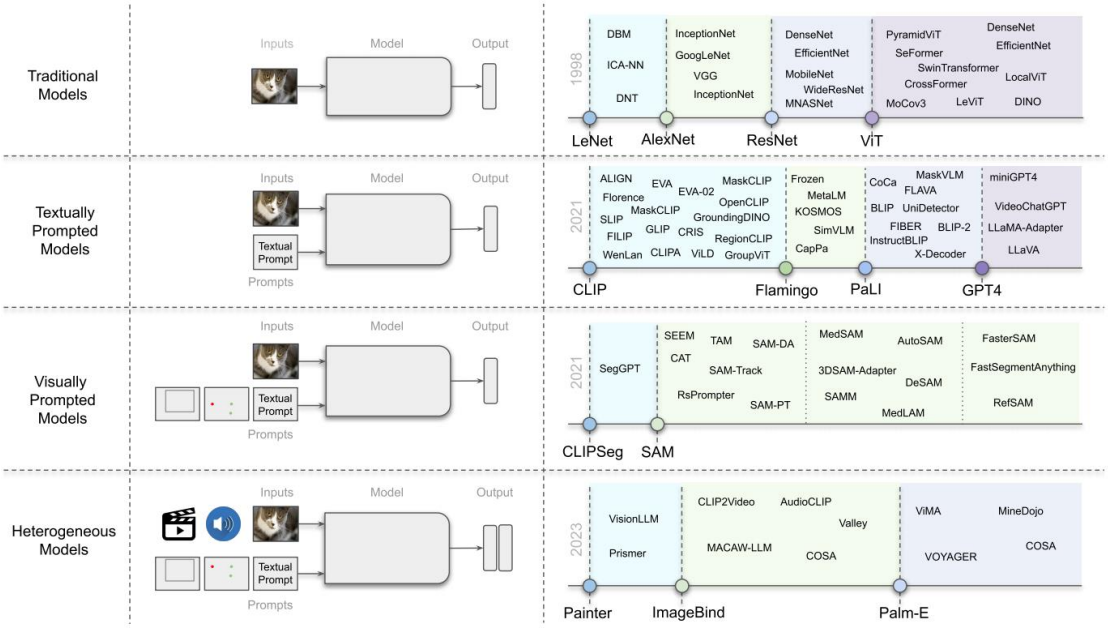

图 1：典型基础模型举例[2]

基础模型的快速发展为计算机视觉的研究带来了全新的动力。近年来（尤其是从 2021 年起），大量研究人员开始着手研究从通用基础模型过渡到下游具体任务的方法，包括目标检测、物体计数、运动估计、数据标注、视频目标跟踪、遥感图像处理、医学图像分割等等。

本文将以 SAM 基础模型为例，分析其在各类下游任务中的应用模式。

2.2 Segment Anything Model（SAM）模型介绍

Segment Anything Model[1]（SAM） 是一个提示型模型，它在 1100 万张图像上训练了超过 10 亿个掩码（mask），实现了强大的零样本泛化。许多研究人员认为 SAM 已经学会了“物体是什么”的一般概念，甚至是未知的物体和不熟悉

的场景，展示了作为计算机视觉基本模型的巨大潜力。

2.2.1 SAM 的任务定义

论文中定义了可提示分割任务（promptable segmentation），也是本文的主要工作，即允许模型根据任何分割提示 prompt（如前景/背景点、粗略的框或掩码、自由形式文本）返回一个有效的分割掩码 mask。下面的图 2(a)展示了多种不同形式的 prompt。

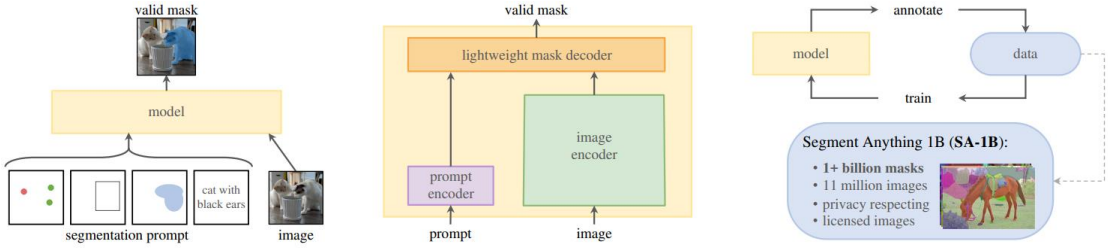

图 2：SAM 模型的任务、架构与数据总体介绍

实际上，对于给定的图片，SAM 生成的 mask 具有不唯一性，在 prompt 较为模糊的情况下，SAM 可能输出不止一个合理的 mask。如图 3 所示，每一列代表同一张图片在给定的 point 条件下生成的不同 mask。

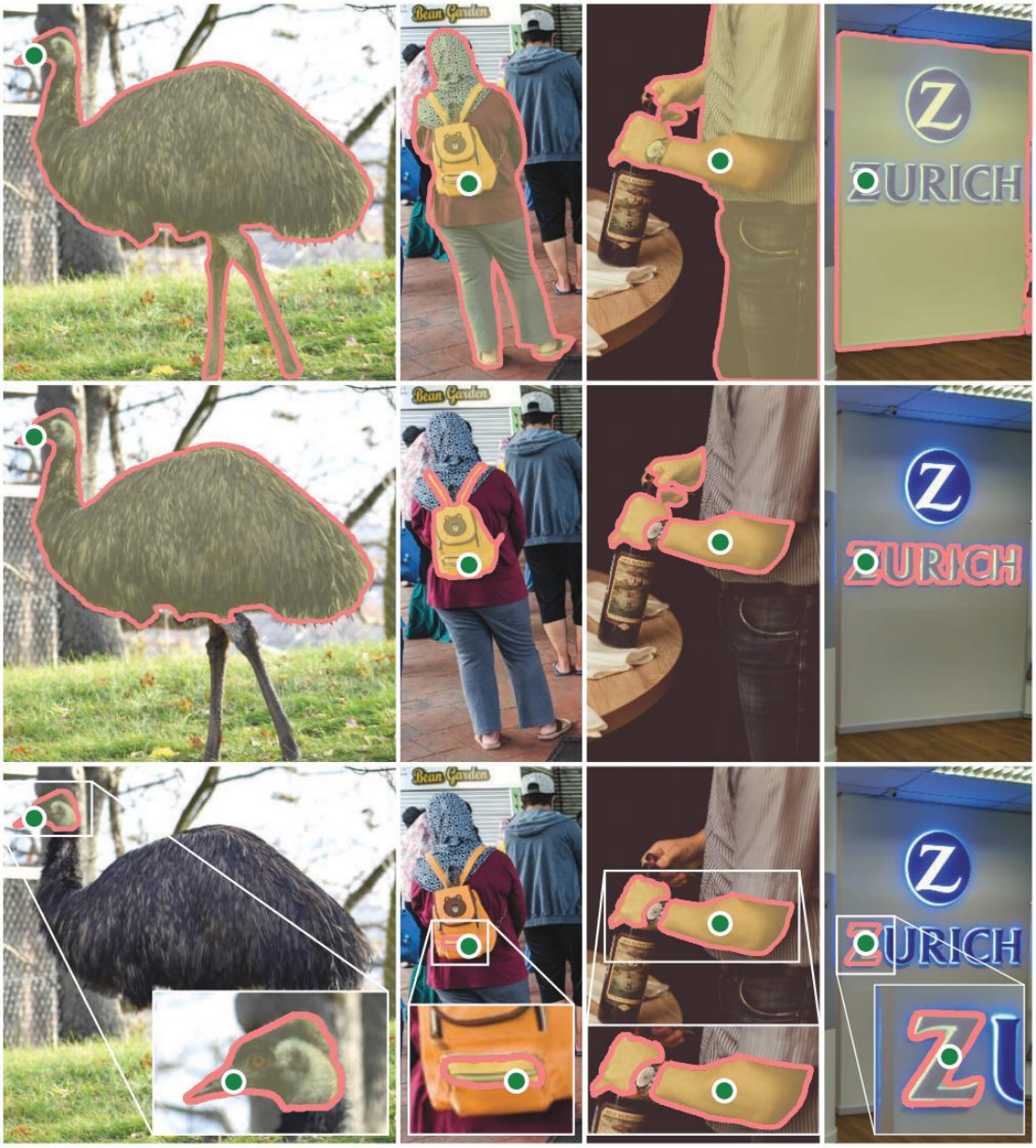

图 3：mask 的不唯一性

2.2.2 SAM 模型架构与训练方法

如图 2(b)所示，SAM 由图像编码器、提示编码器和掩码解码器三部分组成，能实时处理提示并输出掩码。其中：

- 图像编码器（image encoder）是预训练好的 ViT 模型；
- 提示编码器（prompt encoder）处理稀疏和密集输入；
- 掩码解码器（mask decoder）使用自注意力和交叉注意力机制。为了使模型能感知歧义，当 prompt 较为模糊时，模型根据置信度对可能的掩码输出进行

排序，并输出可能性最高的前 n 个 mask。

更具体的模型架构流程图如图 4 所示。当 n 取 3 时，SAM 输出了最右侧的 3 种不同的 mask。

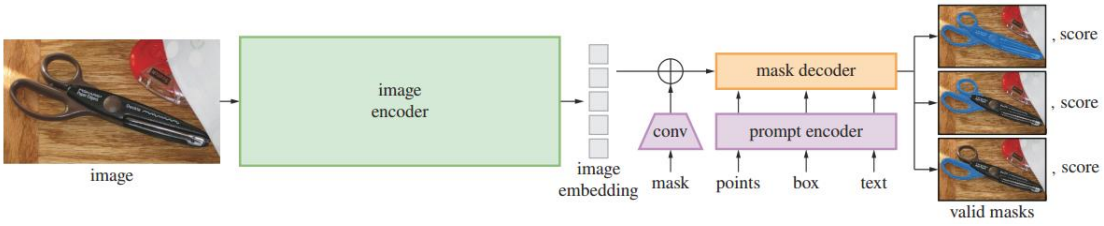

图 4：SAM 模型架构介绍

2.2.3 SAM 的数据引擎

在此之前，学术界还没有专门用于图像分割的大规模数据集。因此该论文建立了一个数据引擎（Data engine），使用 SAM 来协助数据收集，同时利用新收集的数据来改进模型进行迭代。整体的数据构造分为三个阶段：

- 模型辅助的手动标注阶段（Assisted-manual stage）
- 自动预测掩码和模型辅助标注的半自动阶段（Semi-automatic stage）
- 全自动阶段（Fully-automatic stage）

通过以上三步，成功构造了 SA-1B 数据集（可在 https://segment-anything.com/ 上下载），它包含 1100 万张高分辨率的图像和 11 亿个分割掩码。SA-1B 数据集对于不同掩码数量的图片样例如图 5 所示。

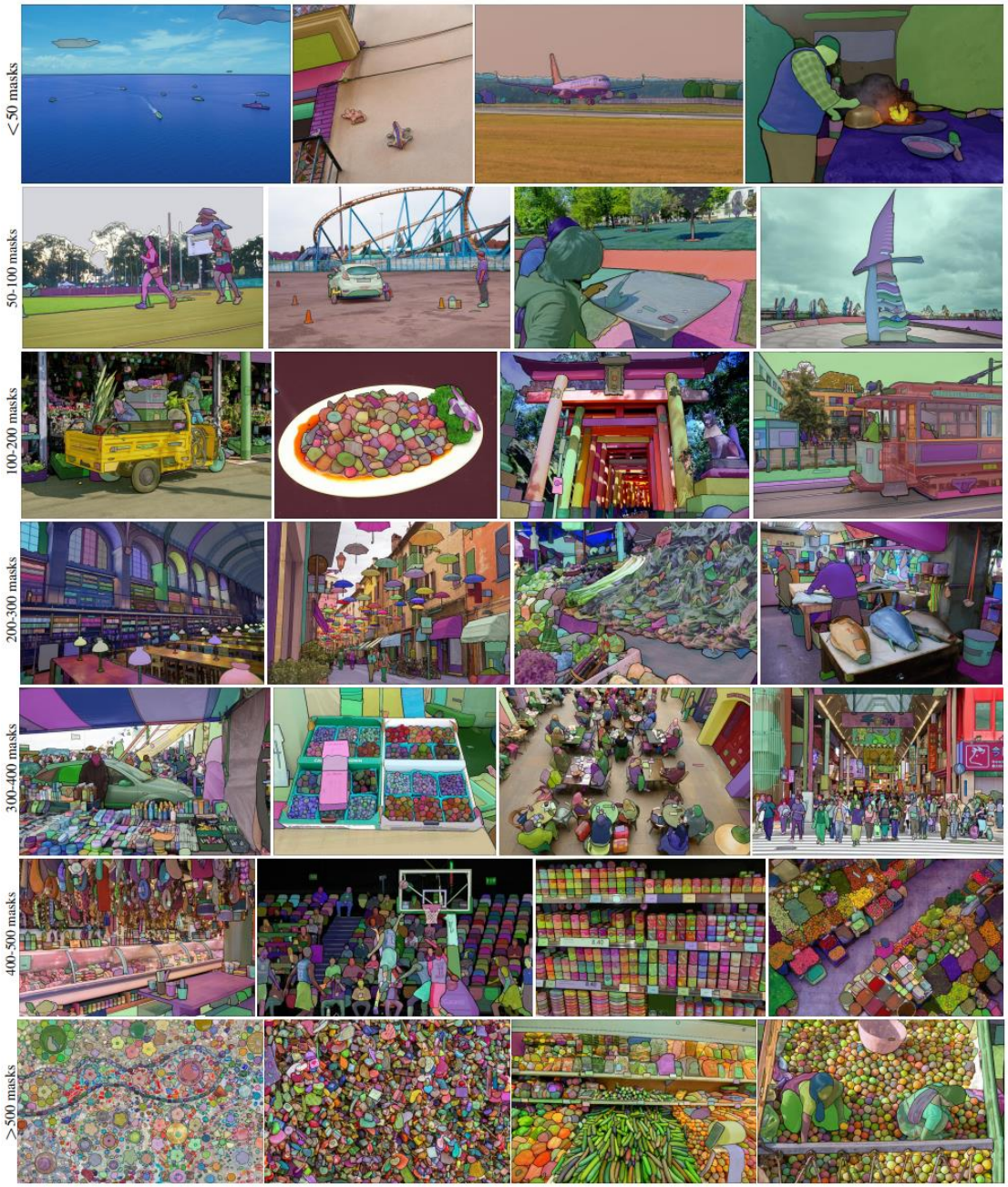

图 5：SA-1B 数据集样例

2.2.4 SAM 模型效果与应用

SAM 在多个零样本下游任务上表现出色，包括零样本单点有效掩码评价、零样本边缘检测、零样本对象建议生成、零样本实例分割和零样本文本转掩码等等。

2.2.5 SAM 模型的局限性

虽然总体而言 SAM 的效果不错，但也并非完美，其缺陷主要体现在：

①SAM 可能会错过一些小尺度结构，在产生 mask 时不一定能做到高度精确。

②当用户给定许多点时，交互式分割方法优于 SAM，因为 SAM 主要是以通用性出发，并非高 IoU 交互式分割思路。

③虽然 SAM 可以完成许多任务，但目前不清楚怎样设计简单的 prompt 来实现语义和全景分割。

④SAM 对文本掩码任务的尝试还是探索性的，并不完全健壮。

3 基于 SAM 模型的衍生技术分析

SAM 模型提出后，产生了大量基于 SAM 的研究，包括图像编辑、风格迁移、复杂场景图像分割、视频目标跟踪、3D 重建等。下面将详细列举相关衍生技术、相应的应用范围和优缺点。

3.1 软件场景（Software Scene）

软件场景需要对图像进行编辑和修复操作，例如移除对象、填充对象和替换对象等等。然而，现有的修复工作都需要对每个掩码进行精细的注释以达到良好的性能。SAM 可以通过简单的提示（如前文所说的点或框）来生成准确的掩码，帮助辅助图像编辑场景。

3.1.1 图像编辑（Image Editing）：Inpaint Anything

Inpaint Anything[3]使用 SOTA 水平的图像修复器（如 LaMa）和 AI 生成内容（AIGC）模型（如 Stable Diffusion），实现了对象移除、对象填充和对象替换的功能。Inpaint Anything 的流程图如图 6 所示。

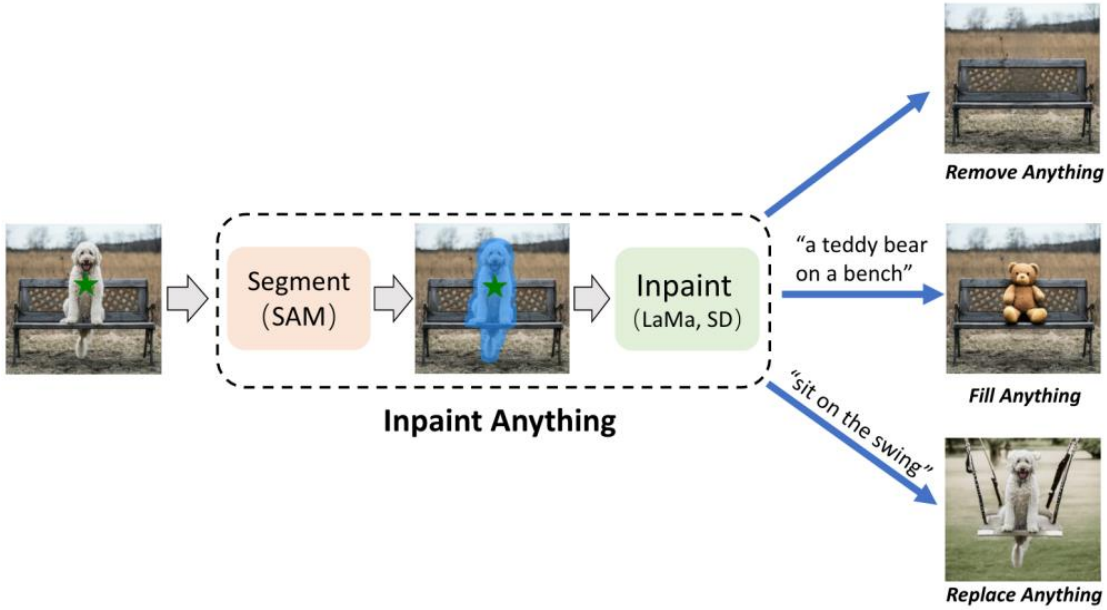

图 6：Inpaint Anything 流程图[3]

算法流程：

对于对象移除（Remove Anything），该流程由 SAM 和图像修复器组成，用

户的点击被视为输入 SAM 的 prompt，从而可以生成指定区域的掩码，然后图像修复器使用 corrosion 和 dilation 操作进行背景的填充，这就完成了对象的移除。

对于对象填充（Fill Anything），第二步会使用 AIGC 模型产生一个符合用户 prompt 的对象，然后将其覆盖原始 mask 即可。

对于对象替换（Replace Anything），与对象填充类似，只不过此时会使用 AIGC 模型产生一个合适的背景，然后将原始 mask 放置在背景的正确位置即可。

3.1.2 图像编辑（Image Editing）：Edit Everything

Edit Everything[4]的流程图如图 7 所示。

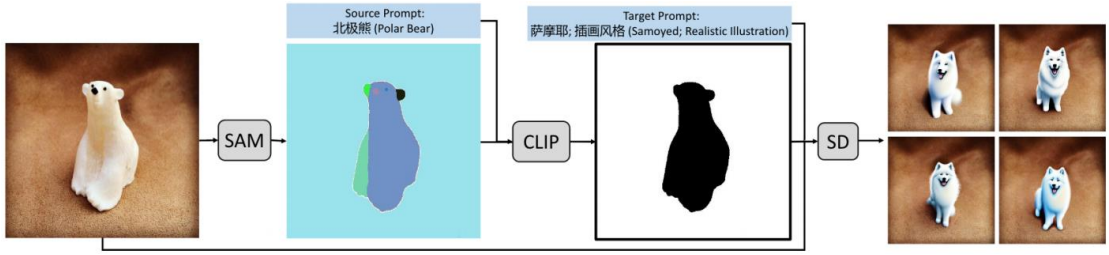

图 7：Edit Everything 流程图[4]

算法流程：

与 Inpaint Anything 类似，SAM 首先在没有提示的情况下将图像分成几个片段，然后依据用户输入的 prompt，使用 CLIP 模型对其进行排序，选择得分最高的片段作为目标段，并使用 Stable Diffusion 生成替换对象。

特点分析：

与 Inpaint Anything 不同的是，Edit Everything 使用了更大规模的模型来处理用户的 prompt，并将复杂的提示分解成更小的实体并逐个替换，提高了图像的真实感。不过，尽管它作为一种新颖的工具表现良好，但在不同的场景下，它仍需要特定的数据增强和强化训练才能获得更好的性能。

3.1.3 风格迁移（Style Transfer）：Any-to-any Style Transfer

风格迁移的目的是将给定风格图像的风格转移到另一个内容图像上。通常，传递的风格主要是由风格图像的整体样式或风格图像的局部颜色和纹理决定，并且只会在内容图像上产生一个结果，缺乏用户交互的灵活性。

为了解决这一问题，Any-to-any Style Transfer[5]通过利用 SAM 的提示区域选

择功能，使用户能灵活地指定在风格迁移过程中选择风格图像的某个区域，以及把该风格应用到某些内容区域。其流程图如图 8 所示。

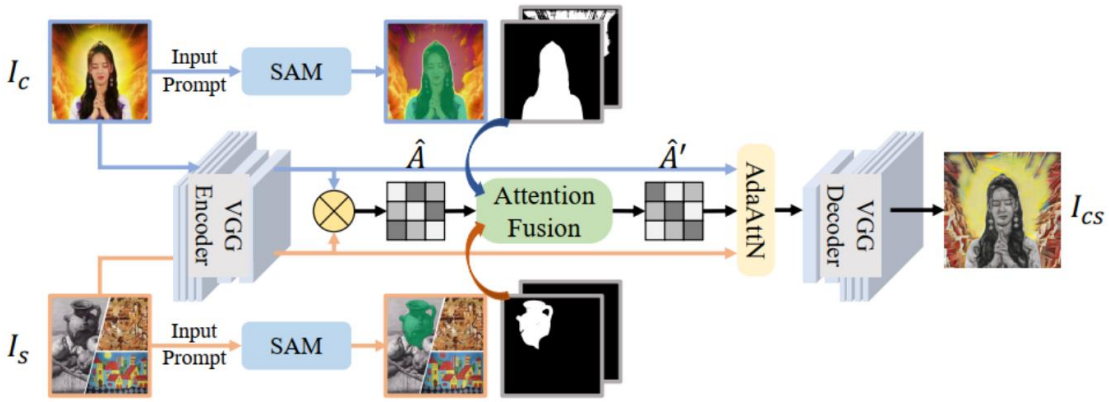

图 8：Any-to-any Style Transfer 流程图[5]

算法流程：

• 使用预训练的 VGG 模型对风格图像和内容图像进行编码，并计算 content-style attention map。

• 通过 SAM 和输入的 prompt 得到风格和内容的掩码。

• 将 attention map 与上一步的掩码信号融合。

• 利用更新后的 attention map 计算风格化特征，并输出最终图片。

特点分析：

这一工作可以极大地方便用户选取自己感兴趣的风格，而排除掉不感兴趣的风格，避免其影响最终输出图片的效果。同时，它可以作为一个插件直接应用在基于局部转换的风格迁移、基于全局转换的风格迁移和基于扩散的风格迁移等多个领域，可见它具有广泛应用的巨大潜力。

3.2 复杂场景（Complex Scene）

除了上述常规场景之外，SAM 还能帮助解决复杂场景（如低对比度场景）中的分割问题。

3.2.1 低对比度场景（Low-contrast Scene）：Segment Any Anomaly+

一个名为 Segment Any Anomaly+[6]（SAA+）的框架可以用于零样本低对比度图像异常分割，其算法流程图如图 9 所示。

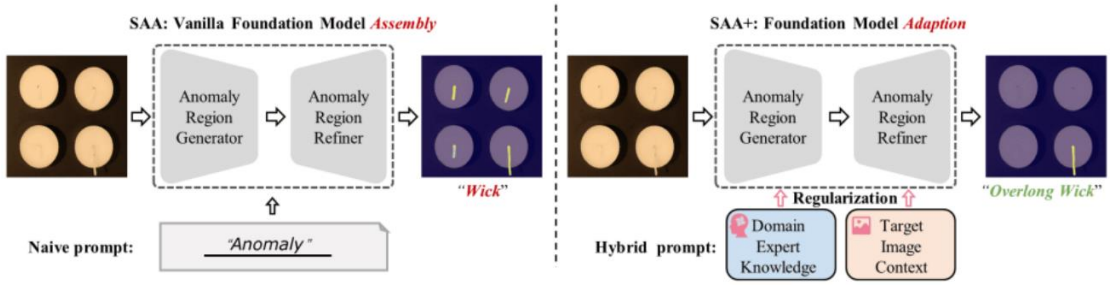

图 9：SAA+流程图[6]以及与 SAA 流程的对比

算法流程：

SAA+与 SAA 框架的不同之处在于，该框架利用混合提示规范化来提高基础模型的适应性，将先验基础知识与目标图像内容进行融合与正则化，从而无需特定领域的微调就能进行更精确的异常分割。

特点分析：

SAA+在 F1-score 像素级分割维度的多个异常分割基准测试中取得了最佳性能，包括 VisA 和 MVTec-AD 数据集。SAA+有效利用了基础模型的能力，并将其调整为零样本异常分割，在检测各种类型的异常方面优于先前的方法。

3.2.2 低对比度场景（Low-contrast Scene）：WS-SAM

针对低对比度图像分割的另一种方法为 WS-SAM[7]，它利用 SAM 进行弱监督的隐蔽物体分割训练，改善了分割与周围环境融为一体的对象的精确度。其算法流程图如图 10 所示。

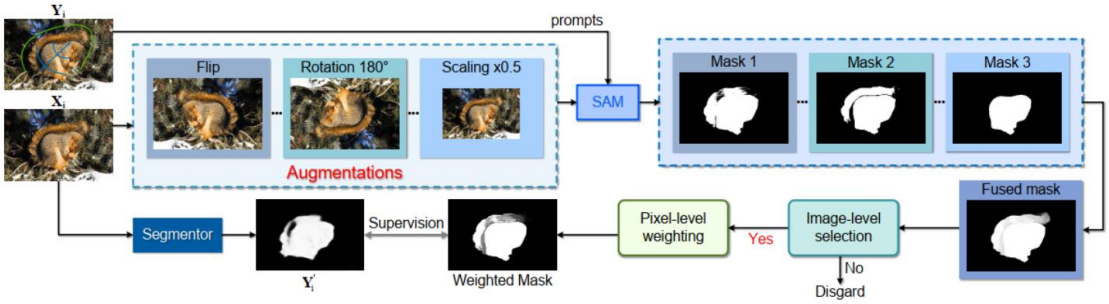

图 10：WS-SAM 流程图[7]

算法流程：

WS-SAM 首先使用 SAM 生成高质量的密集掩码，将稀疏注释作为提示，并将密集掩码用作伪标签来训练分割模型。然后引入了 Multi-scale Feature Grouping（MFG）模块，该模块在不同粒度上对特征进行分组，鼓励分割的一致性，并

为各种隐蔽场景提供完整的分割结果。

特点分析：

由上图可知，WS-SAM 使用了多种数据增强手段，包括翻转、旋转、缩放等。其主要创新点包括基于 SAM 的伪标签和多粒度特征分组，以提高模型学习和区分隐蔽物体与背景的能力。仅仅使用粗糙的监督，SAM 就可以生成足够好的掩码以训练分割器。

3.3 其它技术

3.3.1 视频目标跟踪：Track Anything Model

视频目标跟踪是在视频帧中定位特定目标并随后在整个视频中跟踪它的过程。

Track Anything Model[8]（TAM）主要用于视频中高性能的目标跟踪和分割。与现有方法不同的是，该方法结合了 SAM 并采用交互式方法进行初始化。其流程图如图 11 所示。

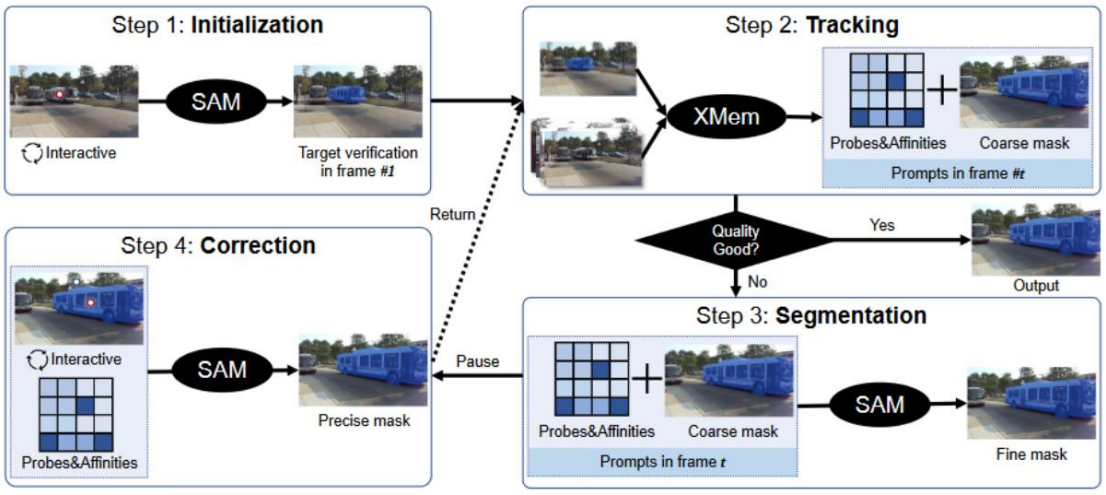

图 11：TAM 流程图[8]

算法流程：

• 使用 SAM 进行初始化。由于 SAM 提供了一个利用弱提示（例如点和边界框）分割感兴趣区域的机会，因此可以使用它给出目标对象的初始掩码。

•使用 XMem 跟踪。在得到初始掩码后，XMem 对随后的帧进行半监督 VOS。当掩码质量不太好时，保存 XMem 的预测和相应的中间参数，即探针和关联度。

• 使用 SAM 进行细化。在 VOS 模型的推断过程中，保持一致且精确地预测

掩码具有很大的挑战性，因此当 XMem 的预测掩码质量不太好时可以使用 SAM 进行细化。

• 人为干预进行修正。经过上述三步，TAM 可以成功解决一些常见的挑战并预测分割掩码，但可能会随着视频时间的逐渐推移而变得不正确。用户可以强制停止 TAM 进程，并使用正面和负面的点击来纠正当前帧的掩码。

特点分析：

TAM 在复杂情况下表现出优异的性能和很强的用户互动性，能够有效地解决视频对象感知中的难题，在许多领域具有潜在的应用前景。

3.3.2 3D 重建：SA3D

SAM 目前仅限于 2D 图像数据，不能直接应用于 3D 场景理解。

在 SAM 的基础上，有研究人员提出了 SA3D[9]。这是一个基于 NeRF 的框架，除了实现精细的 3D 分割外，还可用于 3D 重建，其算法流程图如图 12 所示。

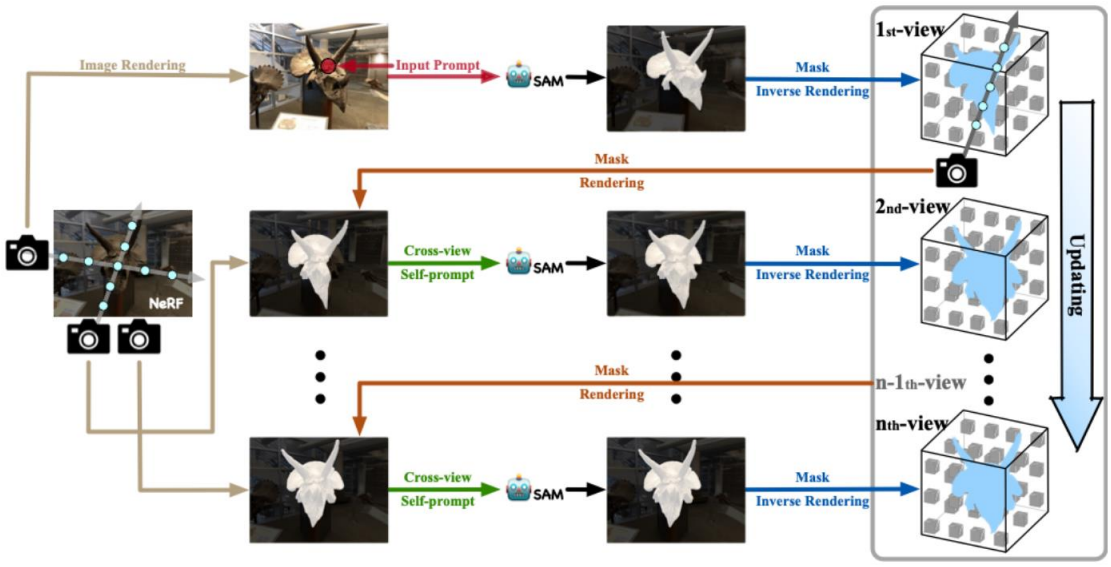

图 12：SA3D 流程图[9]

算法流程：

• 用户为目标对象给出提示，SAM 为它生成一个分段掩码。

•考虑密度分布，利用掩码反渲染技术将该二维掩码投射到三维掩码网格上。

• SA3D 通过从新视图渲染不完整的 2D 掩码，提取可靠的提示并查询 SAM 以获得准确的分割掩码。这个迭代过程被称为交叉视图自提示。

• 在掩码反渲染和获得分割掩码之间交替进行交叉视图自提示过程，最终得

到 3D 分割结果。

特点分析：

与基于 NeRF 的先前方法相比，SA3D 可以在不改变和重新训练任何预训练 NeRF 的情况下轻松适应它们，这依托于 SAM 本身强大的零样本泛化能力。当然，SA3D 还存在一定的缺点。由于 NeRF 方法具有高内存需求和计算复杂度，目前仅适用于相对较小的场景，无法处理大规模的户外场景。

3.4 小结

SAM 模型在图像编辑、风格迁移、复杂场景图像分割、视频目标跟踪、3D 重建等多个领域都可以提供一定的助力。当然，在分析 SAM 模型的意义时，必须要注意不同领域各自面临的问题，辨别不同衍生技术的应用范围，针对性地应用 SAM。

对于技术的发展趋势，我有几点看法。目前多模态预训练大模型的火热可能会为 SAM 注入新的活力，可以考虑结合视觉、语言、声音等多模态信息进行模型训练，增强 SAM 模型对不同场景的适应能力。未来 SAM 模型也应该会往计算效率、鲁棒性和应用范围拓展方向发展，有望在各种下游应用中发挥更大的作用。

4 全文总结

本文首先引出了通用基础模型在计算机视觉研究中的重要性，随后以 Segment Anything Model（SAM）模型为例，介绍了 SAM 模型的思想、架构与性能，并分别列举了 SAM 在图像修复、风格转换、视频目标跟踪等各种任务中的扩展工作和应用。最后，总结了 SAM 在各种图像处理下游任务中的优势、局限性以及对未来研究的展望。

5 参考文献

[1] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W., Dollár, P., & Girshick, R.B. (2023). Segment Anything. ArXiv, abs/2304.02643.

[2] Zhang, C., Liu, L., Cui, Y., Huang, G., Lin, W., Yang, Y., & Hu, Y. (2023). A Comprehensive Survey on Segment Anything Model for Vision and Beyond. ArXiv, abs/2305.08196.

[3] T. Yu, R. Feng, R. Feng, J. Liu, X. Jin, W. Zeng, and Z. Chen, "Inpaint anything: Segment anything meets image inpainting," arXiv preprint arXiv:2304.06790, 2023.

[4] D. Xie, R. Wang, J. Ma, C. Chen, H. Lu, D. Yang, F. Shi, and X. Lin, "Edit everything: A text-guided generative system for images editing," arXiv preprint arXiv:2304.14006, 2023.

[5] S. Liu, J. Ye, and X. Wang, "Any-to-any style transfer: Making picasso and da vinci collaborate," arXiv e-prints, pp. arXiv–2304, 2023

[6] Cao, Y., Xu, X., Sun, C., Cheng, Y., Gao, L., & Shen, W. (2023). 2nd Place Winning Solution for the CVPR2023 Visual Anomaly and Novelty Detection Challenge: Multimodal Prompting for Data-centric Anomaly Detection.

[7] C. He, K. Li, Y. Zhang, G. Xu, L. Tang, Y. Zhang, Z. Guo, and X. Li, "Weakly-supervised concealed object segmentation with sam-based pseudo labeling and multi-scale feature grouping," arXiv preprint arXiv: arXiv:2305.11003, 2023.

[8] J. Yang, M. Gao, Z. Li, S. Gao, F. Wang, and F. Zheng, "Track anything: Segment anything meets videos," arXiv preprint arXiv:2304.11968, 2023.

[9] J. Cen, Z. Zhou, J. Fang, W. Shen, L. Xie, X. Zhang, and Q. Tian, "Segment anything in 3d with nerfs," arXiv preprint arXiv:2304.12308, 2023.